Variant Time-Frequency Cepstral Features for Speaker Recognition

Wei-Qiang Zhang, Yan Deng, Liang He, Jia Liu

Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

wqzhang@tsinghua.edu.cn

Abstract
In speaker recognition (SRE), the commonly used feature vector is basic cepstral coefficients concatenating with their delta and double delta cepstral features. This configuration is borrowed from speech recognition and may be not optimal for SRE. In this paper, we propose a variant time-frequency cepstral (TFC) features, which is based on our previous work for language recognition. The feature vector is obtained by performing a temporal discrete cosine transform (DCT) on the cepstrum matrix and selecting the transformed elements in a specific area with large variances. Different shapes and parameters are tested and the optimal configuration is obtained. Experimental results on the 2008 NIST speaker recognition evaluation short2 telephone-short3 telephone test set show that the proposed variant TFC is more effective than the conventional feature vectors.

Index Terms: Speaker recognition (SRE), time-frequency cepstrum (TFC).

1. Introduction
Speaker recognition (SRE) is the task of recognizing the identity of a speaker. It has many applications such as in information security and identity authentication systems [1, 2]. In recent years, significant improvements have been achieved through different modeling methods, such as Gaussian mixture model – universal background model (GMM-UBM) [3], GMM super-vectors for support vector machines (GSV-SVM) [4], and maximum likelihood linear regression transforms as features for SVM (MLLR-SVM) [5].

In the meanwhile, many different temporal or frequency features have been developed to reflect different aspects of speaker characteristics, including acoustic features, phonotactic features and prosodic features. In addition to the commonly used Mel-frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP) and linear prediction cepstral coefficients (LPCC) [6], there are other novel feature vectors having been presented, such as instantaneous frequencies (IF) [7], linear frequency cepstral coefficients (LFCC) [8], and prosodic features [9].

To encode the temporal information of time-varying spectral features and reduce the effect of channel convolution distortions, delta and double delta (acceleration) features are usually appended to the basic ones to form higher dimensional feature vector. It is a commonly used and successful technology in speech signal processing field. However, this configuration may be not optimal for speaker recognition. On the one hand, the delta vector can be obtained by high-pass linear filtering of the basic vector sequence. In the information theory viewpoint, the information of delta vector is all contained in the basic vector sequence. The simple concatenation is not easy to select the most representative and informative elements. On the other hand, the delta and double delta may bring correlation into the new feature vector, which will depress the performance of some backend classifiers, such as the commonly used diagonal GMM.

In our previous work, we have presented a time-frequency cepstral (TFC) feature [10], which utilizes a temporal discrete cosine transform (DCT) on the cepstrum matrix, and outperforms the widely used shifted delta cepstrum (SDC) for language recognition. In this paper, we will extend this work for speaker recognition.

The remainder of this paper is organized in the following. In section 2, the TFC and variant TFC features are introduced. And in section 3, experimental results are presented to search the optimal parameters and to validate the performance of TFC feature. At last, the conclusion is given in section 4.

2. Variant time-frequency cepstrum
The TFC feature is initially proposed for language recognition [10], which is derived from cepstral-time matrix (CTM) feature [11] and shifted delta cepstrum (SDC) feature [12]. As shown is Figure 1, in the TFC feature extraction, the successive frames of basic feature vectors within a context width window are first extracted to form a cepstrum matrix. A temporal (in horizontal direction) DCT is then performed on the cepstrum matrix and the elements in the upper-left triangular area are in a zigzag scan order. Our motivation is to remove correlation in temporal direction and preserving the elements with greater variability. This is similar to image compression coding, and the detailed discussion can be found in [10].

The procedure of TFC is equivalent to perform a two-dimensional (2D) DCT on spectrum-time matrix. The 2D DCT approach can be interpreted as a compression of the information by a DCT truncation. This process is similar to what is usually done in image compression to reduce the size by using 2D DCT. In addition to the dimension reduction, the compression has two benefits in terms of pattern recognition. Firstly, the classifier can be simpler, since some of the correlation is removed by the temporal DCT. Secondly, the truncation of the higher order vectors helps to reduce the variability due to small scale acoustic events.

In language recognition, the context width is about 20 frames. This leads the variances pattern of the cepstrum matrix after a horizontal DCT is nearly a isosceles triangle, thus we can perform a zigzag scan to select elements in this area to form the TFC feature. In speaker recognition, however, the optimal configuration is not same as in language recognition. As we know, the context width is much less than that in language recognition. Thus the variances pattern becomes narrower.

To give a simple demonstration, the variance of each element of the cepstrum matrix (using successive 9 frames of 13-
dimensional MFCC basic feature vector) after a horizontal DCT was computed using NIST SRE04 1-side training set. The normalized variances (normalized by the maximum elements) are plotted in Figure 2. According to the variance pattern of Figure 2, we can observe that there obviously are other possible configurations besides the isosceles triangle as in TFC, such as a nonsymmetric triangle, trapezoid or even rectangle.

![Figure 1: Procedures of TFC feature extraction [10].](image)

![Figure 2: The normalized variances of each element of the cepstrum matrix after a horizontal DCT.](image)

We will focus on searching the optimal shape and parameters of TFC for speaker recognition through experiments.

### 3. Experiments

#### 3.1. Experimental setup

In the experiments, a classical gender-dependent GMM-UBM system as described in [3] has been built as classifier. A UBM with 256 diagonal mixtures is trained. Speaker models are obtained by maximum a posteriori (MAP) adaptation. In the enrollment stage, only means are adapted, and in the test stage, only top-8 mixtures are used to for log-likelihood ratio scoring.

For the frontend, speech/silence segmentation is performed by a G.723.1 VAD detector [13]. 13-dimensional MFCC coefficients are computed using 25 ms window and 10 ms shift. Cepstral mean subtraction and feature warping [14] with a 3 s window are applied for channel mismatch compensation. After that, 25% of low energy frames are discarded using a dynamic threshold. The UBM is trained using SRE04 1-side training set. All the trial experiments are carried out on NIST SRE08 [15] short2 telephone-short3 telephone test condition.

Both equal error rate (EER) and minimum detection cost function (min DCF) are used as the performance measure.

#### 3.2. Baseline performance

We first use the 13-dimensional MFCC with delta and double delta as feature vector. The performance of female and male conditions are listed in Table 1. The EERs are about 12% and the min DCFs are about $6 \times 10^{-2}$.

#### 3.3. Parameter optimization

In TFC feature, the context width (the width of cepstrum matrix in Figure 1 and the shape of reserved area after DCT are the changeable parameters. From Figure 2, we can observe that the main variance is roughly located in $13 \times 3$ rectangle. So we first fix the area, and vary the context width from 7 to 13 with step size 2. The results on are illustrated in Table 2.

From the results, we can see that context width = 9 gives the best performance. (When the EERs and the min DCFs are not consistent, we only consider the min DCFs in this paper.) In MFCC with delta and double delta feature vectors, the context width is also 9, which implies that the discriminative information for different speakers may primarily lie in this context width. The results also show that the optimal context width for speaker recognition is much narrower than that for language recognition, so we must reconsider the parameters.

<table>
<thead>
<tr>
<th>Context width</th>
<th>Gender</th>
<th>EER(%)</th>
<th>min DCF $\times 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>female</td>
<td>12.31</td>
<td>5.61</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>9.73</td>
<td>4.57</td>
</tr>
<tr>
<td>9</td>
<td>female</td>
<td>12.20</td>
<td>5.64</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>9.38</td>
<td>4.37</td>
</tr>
<tr>
<td>11</td>
<td>female</td>
<td>11.78</td>
<td>5.67</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>9.22</td>
<td>4.46</td>
</tr>
<tr>
<td>13</td>
<td>female</td>
<td>11.81</td>
<td>5.65</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>9.26</td>
<td>4.55</td>
</tr>
</tbody>
</table>

**Table 1:** Baseline performance with MFCC $13+\Delta+\Delta\Delta$ feature vectors, SRE08 short2 telephone-short3 telephone.

<table>
<thead>
<tr>
<th>Feature vector</th>
<th>Gender</th>
<th>EER(%)</th>
<th>min DCF $\times 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC $13+\Delta+\Delta\Delta$</td>
<td>female</td>
<td>13.29</td>
<td>6.05</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>10.46</td>
<td>5.22</td>
</tr>
</tbody>
</table>

**Table 2:** Comparison of variant TFC with different context width, rectangle $13 \times 3$, SRE08 short2 telephone-short3 telephone.
After obtaining the optimal context width, we will find the optimal shape of the reserved area. We first approximate the area as a rectangle. In order to give fair comparison, we fix the total dimensions roughly.

The rectangle is changed from 19×2 to 13×3 and to 10×4, and the results are shown in Table 3. From this table, we can observe that when the rectangle is 13×3, the TFC feature vector obtains the best performance.

As we can see from Figure 2, the shape of elements with large variances is not a regular rectangle, and perhaps the trapezoid will be more suitable. Our next experiment is to detail the optimal shape. We tilt the bottom base of the rectangle to form a trapezoid. The slope of the hypotenuse is changed from 0 to 9, and from 13×3 to 15+13+11 to 17+13+9, while keeping the total dimensions as 39. The results are given in Table 4. From Table 4, we can see that the trapezoid shape with 15+13+11 performs the best.

Through the above experiments, we obtained the optimal parameters of TFC feature vector for speaker recognition. The optimal context width is 9 and the optimal shape is trapezoid with three columns 15+13+11. Compared this parameter with the baseline MFCC 13+Δ+ΔΔ, the min DCFs for female and male condition are decreased from 6.05×10^{-2} to 5.60×10^{-2} and from 5.22×10^{-2} to 4.36×10^{-2}, respectively.

Note that we fixed the total dimension as 39 in our experiments. If we relax this constraint, maybe more gains can be achieved with additional computational cost.

### 3.4. Further experiments with channel compensation

To further validate the TFC features, we apply the feature domain latent factor analysis (fLFA) [16] to the MFCC and TFC feature vectors. In fLFA, a gender-independent UBM with 512 mixtures is used and the channel number is set as 30. The fLFA training data come from SRE04 8-side training set. After fLFA, the gender-dependent UBM with 2048 mixtures are retrained. At last, zt-norm [17] is performed on the raw log-likelihood ratio scores. The zt-norm data come from SRE05 and SRE06. The zt-norm model number for female and male are 339 and 239 respectively and t-norm segment number for female and male are 199 and 274 respectively.

The results with fLFA and zt-norm are listed in Table 5 and the detection error trade-off (DET) curves are plotted is Figure 3. From the results, we can see that consistent improvements can be obtained when the feature-domain channel compensation and back-end score normalization are both used. The min DCFs for female and male condition are decreased from 5.22×10^{-2} to 4.08×10^{-2} and from 3.97×10^{-2} to 3.11×10^{-2}, respectively.

![Figure 3: The DET curves of MFCC 13+Δ+ΔΔ and TFC 15+13+11 with fLFA and zt-norm, SRE08 short2 telephone-short3 telephone. The circles denote the min DCF operating points.](image)

In addition to the traditional DCF, we are also interested in the new DCF defined in NIST SRE 2010 [18]. In the new DCF, the target prior is set as 0.001, which requires much more non-target trials to give meaningful measurement. To do this, we extend the SRE08 trial lists by performing a Cartesian product of

<table>
<thead>
<tr>
<th>Rectangle</th>
<th>Gender</th>
<th>EER(%)</th>
<th>min DCF×100</th>
</tr>
</thead>
<tbody>
<tr>
<td>19×2</td>
<td>female</td>
<td>14.30</td>
<td>6.28</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>9.34</td>
<td>4.61</td>
</tr>
<tr>
<td>13×3</td>
<td>female</td>
<td>12.20</td>
<td>5.64</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>9.38</td>
<td>4.37</td>
</tr>
<tr>
<td>10×4</td>
<td>female</td>
<td>12.39</td>
<td>5.97</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>9.89</td>
<td>4.78</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trapezoid</th>
<th>Gender</th>
<th>EER(%)</th>
<th>min DCF×100</th>
</tr>
</thead>
<tbody>
<tr>
<td>13×3</td>
<td>female</td>
<td>12.20</td>
<td>5.64</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>9.38</td>
<td>4.37</td>
</tr>
<tr>
<td>15+13+11</td>
<td>female</td>
<td>11.96</td>
<td>5.60</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>9.37</td>
<td>4.36</td>
</tr>
<tr>
<td>17+13+9</td>
<td>female</td>
<td>13.23</td>
<td>5.98</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>8.86</td>
<td>4.48</td>
</tr>
</tbody>
</table>
model set and test segment set. We use Brummer’s toolbox [19] to plot NIST’s SRE min DCF as a function of the target prior. The results are showed in Figure 4. We can see that by substituting TFC 15+13+11 for MFCC 13+Δ+ΔΔ, the normalized min DCFs for female and male condition are decreased from 0.758 to 0.698 and from 0.611 to 0.573, respectively, which shows the effectiveness of TFC feature vector.

Figure 4: The normalized min DCF curves of MFCC 13+Δ+ΔΔ and TFC 15+13+11 with fLF A and zt-norm, SRE08 short2 telephone-short3 telephone Cartesian test. The DCF is normalized by \( \min(P_{tar}, 1 - P_{tar}) \).

### 4. Conclusions

In this paper, we have proposed the variant time-frequency cepstral features for speaker recognition. This is an extension of our previous work of TFC. Through analyzing the variance pattern of cepstrum matrix after horizontal DCT transform, we find the area with greater variance is not an isosceles triangle but nearly a rectangle. Coupled with detailed experiments, we obtain the optimal configuration is a trapezoidal shape with context width 9 and three columns 15+13+11. Experiments on NIST SRE08 short2 telephone-short3 telephone test set show that TFC is more effective than MFCC.

### 5. Acknowledgements

This work was supported by the National Natural Science Foundation of China and Microsoft Research Asia under Grant No. 60776800, by the National Natural Science Foundation of China under Grant No. 90920302, and in part by the National High Technology Development Program of China (863 Program) under Grant No. 2006AA010101, No. 2007AA04Z223, No. 2008AA02Z414 and No. 2008AA040201.

### 6. References


